Emerging Business Applications of High Performance Analytics

August 2014

Tan Yaw, Sr. Data Scientist
Table of Contents

• Introduction
• Data Lake
• Analytics
• Labs
• **New Independent Venture:** Spun out & jointly owned by EMC & VMware
• **Top Talent:** 1700~ employees
• **Proven Leadership:** Paul Maritz, CEO
• **Global Customer Validation:** +1000 Tier-1 Enterprise Customers
• **Strategic Backing:** $105M investment by GE
• **Bold Vision:** New platform for a new era, focused on the intersection of Big Data, PaaS, and Agile Software Development
EMC Federation
Pivotal – What we do

Pivotal Data Labs

\[ f(X) = \frac{1}{n} \sum_{i=1}^{n} T_i(X) = \frac{1}{n} \sum_{m=1}^{M} \sum_{i=1}^{n} W_m(X) Y_i = \frac{1}{n} \left( \sum_{m=1}^{M} W_m(X) \right) Y_i \]

SAS Event: Kuala Lumpur - Hadoop, Big Data & Analytics

© Copyright 2013 Pivotal. All rights reserved.
Customer Reference

- Groupon
- LinkedIn
- Twitter
- American Express
- Reuters
- Facebook
- China Citic Bank
- RBC Royal Bank
- NYSE Euronext
- SBI

SAS Event: Kuala Lumpur - Hadoop, Big Data & Analytics
Data Lake
Big Data

- Big Data is not singularly about ‘large size’
Pivotal Business Data Lake Architecture

Unified Sources
- Real-time ingestion
- Micro batch ingestion
- Batch ingestion

Centralized Management
- System monitoring
- System management

Unified Data Management Tier
- Data mgmt. services
- MDM
- RDM
- Audit and policy mgmt.

Workflow Management
- Processing Tier
  - In-memory
  - MPP database
- Distillation Tier
  - HDFS storage
  - Unstructured and structured data

Flexible Actions
- Real-time insights
- Interactive insights
- Batch insights

SAS Event: Kuala Lumpur - Hadoop, Big Data & Analytics
Emerging Analytics Architecture

Data Ingestion
Streams/Feeds

Data Staging Platform

Predictive Analytics
Data Science

Descriptive Analytics
Business Analysis

Analytic Data Marts
MPP Database

Enterprise Data Warehouse
RDBMS

SAS Event: Kuala Lumpur - Hadoop, Big Data & Analytics

© Copyright 2013 Pivotal. All rights reserved.
## How is Business Data Lake Different

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Business Data Lake</th>
<th>EDW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Data Model</td>
<td>Single Standard Data view = Base class</td>
<td>Single Class = Single view across the enterprise</td>
</tr>
<tr>
<td></td>
<td>Enhanced Local Data view = Derived Classes</td>
<td></td>
</tr>
<tr>
<td>Data Quality</td>
<td><img src="image" alt="Full Spectrum" /></td>
<td><img src="image" alt="Clean" /></td>
</tr>
<tr>
<td>Data Integration</td>
<td>SQL, SAS, R, MapReduce, NoSQL</td>
<td>SQL access &amp; Integration with SAS, R, ...</td>
</tr>
<tr>
<td>Multiple Interfaces</td>
<td>Mixed workload with varying QoS</td>
<td>Limited QoS separation required</td>
</tr>
<tr>
<td>Quality of Service</td>
<td><img src="image" alt="Low Latency" /> - Interactive - Batch</td>
<td></td>
</tr>
</tbody>
</table>
Hadoop at the Center

- CLOUD FOUNDRY
  Hadoop as a Service

- HAWQ
  Fastest SQL Query Engine

- PIVOTAL
  Enabling the Data Driven Enterprise

- Spring XD
  Building Big Data Apps

- GemFire XD
  In-Memory Real-time Analytics

- PIVOTAL DATA DISPATCH
  Big Data On-Demand

SAS Event: Kuala Lumpur - Hadoop, Big Data & Analytics

© Copyright 2013 Pivotal. All rights reserved.
China Citic Bank implements Data Lake to integrate multiple databases and in-database modelling for rapid model deployment.

Opportunity: Integrate the bank’s FICO TRIAD Customer Management Solution, Database Marketing platform, IBM Cognos Business Intelligence software, and subcenter customer relationship management (CRM)

Business Benefits:

• More Productive Telephone Sales Center
• Optimized Marketing Campaigns (1286 with 86% reduction in configuration time)
• Faster model deployment via ‘In-database’ analytics
The NYSE Euronext manages exponential data growth and supports analytic applications.

“Pivotal rings the NYSE bell on Oct 29”

Opportunity: Work with NYSE technologies division on new historical archive platform.

Solution: Data Infrastructure to allow NYSE to handle trading data real time.

Co-developed in partnership with NYSE Technologies, Pivotal Data Dispatch is aimed squarely at the big data information worker. The idea of this product is to provide data analysts with an easy way to provision various big data sets from any source, including Hadoop, MPP, flat files or legacy databases.
Improving security analytics and implementing Data Lake architecture based on Pivotal HD, HAWQ and GPDB

**Business Challenge:** Improving security analytics for credit card transactions and developing “Data Lake” architecture for future projects.

**Volume:** 4 TBs and growing

**Solution:** Data Lake architecture including Pivotal HD, HAWQ and
Analytics
Moneyball

- Lost your best players
- No resources
- Competing against richer, better opposition

Q: How do you compete?
A: Data Driven Analytics
Man vs. Machine: Simple Charts

- Traditionally, ‘Man’ takes data and turn them into charts in order to visualize relationships. Charts are simple and easy to interpret.
- ‘Man’ has ‘Analytical Limits’. We inherently view the world in 2-Dimensions and in simple linear relationships.
Man vs. Machine: Complex relationships

• But the real-world is complex!

• It is not just X and Y relationships. $X \rightarrow Y \rightarrow Z \rightarrow A \rightarrow B \rightarrow C$

• It is not just linear.

• Charts that try and visualize complex relationships are themselves more complex.
Man vs Machine: Finding Patterns

- How do we classify and identify different groups within a dataset
Man vs Machine: Machine Learning

- Machines are able to analyze complex patterns within the data that the human mind has difficulty visualizing.
Evidence-Based Decisions

When somebody on staff asks what we should do to address a problem, the first questions I now ask are ‘What does the research say? What is the evidence base?'

The core idea is that decisions supported by hard facts and sound analysis are likely to be better than decisions made on the basis of instinct, folklore or informal anecdotal evidence.
Decision-Based Evidence

Many managers think they’ve committed their organizations to evidence-based decision making — but have instead, without realizing it, committed to decision-based evidence creation.

When asking staff to conduct a major analysis, a project team told us, “The executives have already made up their minds…. We are being told that this is the way that we are going, we need to get on board and make the decision work out to be [the new choice].”
Route Optimization

Customer
A major courier delivery services company

Business Problem
Optimizing routing decisions while meeting the demand and satisfying the many business constraints to guarantee feasibility and compliance.

Challenges
• Routing problems are known to be NP-Hard
• Size of the operation. Delivery of 3 million packages a day with the largest fleet in the US
• Existing solution takes weeks to roll out monthly routing plans

Solution
• Avoided expensive data movements by addressing demand forecasting and route optimization in the database
• Built a fully parallelized approximation algorithm that featured a variation of Floyd Warshall all pairs shortest paths and neighborhood searches
• Achieved significant reduction in fuel consumption over a greedy initial feasible solution
Predicting Commodity Futures through Twitter

Customer
A major a agri-business cooperative

Business Problem
Predict price of commodity futures through Twitter

Challenges
• Language on Twitter does not adhere to rules of grammar and has poor structure
• No domain specific label corpus of tweet sentiment – problem is semi-supervised

Solution
• Built Sentiment Analysis and Text Regression algorithms to predict commodity futures from Tweets
• Established the foundation for blending the structured data (market fundamentals) with unstructured data (tweets)
Credit Risk Assessment and Stress Testing

**Customer**

A global financial services provider

**Business Problem**

Speed up the process of compliance reporting and stress testing for Basel III.

**Challenges**

Running the calculation procedures on the customer’s legacy database were time-consuming, therefore had to be done in overnight batch mode.

**Solution**

- Implement risk asset calculation and stress testing on the Greenplum database.
- Three years of data was processed in well under 2 minutes, significantly faster than the customer’s current procedures.
- Connect an “in-database” visualization tool to the Greenplum database via ODBC for on-demand reporting and visualization.
Text Analytics for Churn Prediction

Customer
A major telecom company

Business Problem
Reducing churn through more accurate models

Challenges
• Existing models only used structured features
• Call center memos had poor structure and had lots of typos

Solution
• Built sentiment analysis models to predict churn and topic models to understand topics of conversation in call center memos
• Achieved 16% improve in ROC curve for Churn Prediction
Cross-Channel Customer Engagement

Customer
A major health insurance company

Business Problem
As each call to the call center represents a significant cost to the company, find out when customers are using the call center instead of the website

Challenges
- Unstructured text data requires considerable preprocessing

Solution
- Used logistic regression to predict whether a customer would be unable to find their information on the web and need to call in
- Created a topic model based on the call logs to learn what these customers were calling about, since these would be the topics they were having trouble finding on the website
Labs
Industrial-era business practices

- Many enterprise-grade business practices are well suited for an industrial era.
- But may face challenges when dealing with the Internet era where ‘Speed’ and ‘Innovation’ are being key competitive levers.
Industrial-era business practices

- Waterfall Project Mgmt
- Develop, Test, Production, DR environments
- Detailed Requirements
- Structured Data Schema
Knowledge-era business practices

Silicon Valley has always been a hot-bed of innovation.

When working with new technology, demanding high-availability, rapid speed, uncertain customer preferences, DIFFERENT BUSINESS PROCESS are needed.
Industrial Era

- To innovate, to be competitive,
- Stop running Large Project
- Start running Innovation Labs

Knowledge Era

- To deal with a VUCA (Volatile, Uncertain, Complex, Ambiguous) World
- We need Iterations & Learning
Labs Experiments

- Data Lab experiments as a key approach in generating value
Data Labs

Data Engineering + Data Science
MAD Approach to Analytics

**Magnetic** - *attracting* data to your EDW by removing “barriers to entry”

**Agile** – enabling *rapid analyses* through the availability of powerful tools as close as possible to the data

**Deep** – going beyond basic data operations to empower analysts to *reach new, rich depths* in their insights
Conclusion
Analytics Vision

Store Everything
- Obsessively collect data
- Keep it forever
- Put the data in one place

Analyze Anything
- Cleanse, organize, and manage your data lake
- Make the right tools available
- Use the resources wisely to compute, analyze, and understand data

Build the Right Thing
- Use insights to iteratively improve your product
Thank You